Biochimie 186 (2021) 13–27

Contents lists available at ScienceDirect

Biochimie

journal homepage: www.elsevier.com/locate/biochi

Analyses of viral genomes for G-quadruplex forming sequences reveal their correlation with the type of infection

Natália Bohálová [a, b, 1], Alessio Cantara [a, b, 1], Martin Bartas [c], Patrik Kaura [d], Jiří Šťastný [d, e], Petr Pečinka [c], Miroslav Fojta [a], Jean-Louis Mergny [a, f], Václav Brázda [a, *]

[a] Institute of Biophysics of the Czech Academy of Sciences, Královopolská 135, Brno, 612 65, Czech Republic
[b] Department of Experimental Biology, Faculty of Science, Masaryk University, Kamenice 5, 62500, Brno, Czech Republic
[c] Department of Biology and Ecology/Institute of Environmental Technologies, Faculty of Science, University of Ostrava, Ostrava, 710 00, Czech Republic
[d] Brno University of Technology, Faculty of Mechanical Engineering, Technická 2896/2, 616 69, Brno, Czech Republic
[e] Department of Informatics, Mendel University in Brno, Zemědělská 1, Brno, 613 00, Czech Republic
[f] Laboratoire d'Optique et Biosciences, Ecole Polytechnique, CNRS, INSERM, Institut Polytechnique de Paris, 91128, Palaiseau, France

ARTICLE INFO

Article history:
Received 18 January 2021
Received in revised form
30 March 2021
Accepted 31 March 2021
Available online 9 April 2021

Keywords:
G-quadruplex
Viral genome
Bioinformatics
Persistent infection
Acute infection
G4Hunter

ABSTRACT

G-quadruplexes contribute to the regulation of key molecular processes. Their utilization for antiviral therapy is an emerging field of contemporary research. Here we present comprehensive analyses of the presence and localization of putative G-quadruplex forming sequences (PQS) in all viral genomes currently available in the NCBI database (including subviral agents). The G4Hunter algorithm was applied to a pool of 11,000 accessible viral genomes representing 350 Mbp in total. PQS frequencies differ across evolutionary groups of viruses, and are enriched in repeats, replication origins, 5′UTRs and 3′UTRs. Importantly, PQS presence and localization is connected to viral lifecycles and corresponds to the type of viral infection rather than to nucleic acid type; while viruses routinely causing persistent infections in Metazoa hosts are enriched for PQS, viruses causing acute infections are significantly depleted for PQS. The unique localization of PQS identifies the importance of G-quadruplex-based regulation of viral replication and life cycle, providing a tool for potential therapeutic targeting.

1. Introduction

G-quadruplexes (G4s) are stacked secondary nucleic acid structures formed by G-rich DNA or RNA sequences. The building block of a G4 is a G-quartet that arises via Hoogsteen base pairing of four guanines. Two or more G-quartets are stacked on top of each other and physiologically stabilized mainly by monovalent cations, such as potassium [1]. G4s have been found in various genome locations such as telomeres, origins of replication and promoter regions [2,3]. With respect to their presence in these particular areas of genomes, G4s are most likely involved in different cellular events such as modulation of telomeric functions or regulation of gene expression [2,4].

In recent years, increasing evidence showed that G4s are also involved in the control of key viral processes, such as replication or transcription and influence viral latency [5]. Stabilization of G4 by specific G4 ligands, such as bisquinolinium derivatives (PhenDC3), porphyrin derivatives, perylenes and naphthalene diimides (PIPER), pyridostatin (PDS) or acridine derivatives (*e.g.*, BRACO-19) can modulate viral gene expression and, moreover, viral infection [6]. Therefore, G4 ligands can be considered as promising anti-proliferative agents [7] and also as potential antiviral therapeutic compounds [6].

Under physiological conditions, G4 formation is regulated by G4-binding proteins [8]. This suggests their crucial roles in a number of signaling pathways, including those relevant for viral infection (*e.g.*, the immune response). For example nucleolin, the most abundant nucleolar phosphoprotein known to promote the formation of G4 and inhibit the activity of the *c-Myc* gene [9], has also been reported to stabilize the formation of G4 in LTR repeats of the human immunodeficiency virus 1 (HIV-1) virus and thus silence viral transcription [10]. Nucleolin also mediates immune evasion of Epstein-Barr virus (EBV) by stabilizing G4 in EBV-encoded nuclear antigen-1 (EBNA1) mRNA [11] and suppresses replication of Hepatitis C Virus (HCV) by binding to RNA G4 [12].

* Corresponding author.
E-mail address: vaclav@ibp.cz (V. Brázda).
[1] These authors contributed equally.

https://doi.org/10.1016/j.biochi.2021.03.017

Furthermore, viral G4-binding proteins are crucial for the viral life cycle; for example nucleocapsid proteins of the HIV-1 virus can function as molecular chaperones for G4 [13] and it the G4-binding macrodomain within the SARS unique domain (SUD) of SARS-CoV Nsp3 protein is essential for viral replication and transcription [14]. Targeting of viral G4 binding proteins is thus a prospective antiproliferative and antiviral therapeutic strategy. For this purpose, antiviral DNA aptamers have been designed [15].

To assess the propensity of a sequence to form G4 and distribution of these putative G4-forming sequences (PQS) in genomes, several algorithms have already been described. Chronologically, the first generation algorithm searches for the [GnNmGnNoGnNpGn] motif in a sequence [16]. The second generation algorithm takes into consideration occurrence of repeating units of Gn ($n \geq 2$) [17]. Nevertheless, both algorithms produce binary (yes/no; match/no match) results, rather than the qualitative analyses that are necessary for correlation of G4 strength metrics. G4Hunter was developed to overcome this limitation, and it calculates G4 propensity depending on the asymmetrical distribution of G and C (GC-skewness) [18].

Our aim was to analyze for the first time a pool of all accessible 11,000 viral genomes searching for the presence of PQS with the G4Hunter web application [19]. Viruses are small infectious agents that replicate inside a living cell of an organism. According to the Baltimore classification they can be divided into classes according to their genome: double-stranded DNA (dsDNA) viruses, single-stranded DNA (ssDNA) viruses, double-stranded RNA (dsRNA) viruses, positive sense RNA (ssRNA+) viruses, negative stranded RNA (ssRNA-) viruses, RNA reverse transcribing viruses and DNA reverse transcribing viruses. In addition, there are also subviral agents such as viroids and satellites. Viroids are small circular ssRNA molecules, which do not encode any protein. All known viroids infect higher plants and represent dangerous agricultural pathogens [20]. Satellites are also mainly found in plants; their replication depends on co-infection with an independent helper virus [21]. Evidence of the presence of PQS has already been reported in several viruses from different families; for example dsDNA viruses from *Herpesviridae* or *Papillomaviridae*, ssRNA+ viruses from *Flaviviridae* or *Filoviridae*, RNA viruses with reverse transcriptase (RT) *Retroviridae* and also dsDNA with RT *Hepadnaviridae* [6,22]. The presence of PQS and G4 functions in viruses remains poorly understood and until now has been mainly limited to studies of eukaryotic viruses by first generation algorithms [23,24]. In our study, we comprehensively analyzed the presence and locations of PQS in all accessible viral genomes with diverse hosts using G4Hunter. For classification purposes, we categorized viruses according to their types of genome and current family assignment in the National Center for Biotechnology Information (NCBI) database. These data bring basic information about evolutionary changes of PQS frequency among the phyla and for the first time link G4 with the type of viral infection and thus provide possible new strategies for their targeting in therapeutics.

2. Materials and methods

2.1. Selection of viral sequences

All 13,809 accessible viral genomic sequences of 11,000 unique species were downloaded from the genome database of the NCBI on the 10th of January 2020. The phylogenetic classification was performed according to NCBI and Virus-Host database [25]. The accession codes and phylogenetic classification of all viral complete genomic sequences are shown in Supplementary Material 1.

2.2. Process of analysis

We used the computational core of our DNA analyzer software written in Java with G4Hunter algorithm implementation [19]. Parameters for G4Hunter were set to 25 nucleotides for window size and a threshold score of 1.2 or above. A separate list of PQS in each of the sequences and an overall report were obtained. The overall results for each species group contained a list of species with size of genomic DNA, GC content, number of PQS found, frequency normalized for 1,000 nt and length of the sequence covered with PQS (Supplementary Material 2). PQS were also grouped depending on their G4Hunter score in the five following intervals: 1.2–1.4, 1.4–1.6, 1.6–1.8, 1.8–2.0 and 2.0 or more (the higher the score, the more likely/more stable is the G4).

2.3. Statistical analysis

Statistical evaluations of differences in PQS between phylogenetic groups, viruses grouped by the type of infection and features were made by Kruskal–Wallis test. Post-hoc multiple pairwise comparison by Dunn's test with Bonferroni correction of the significance level was applied with *p*-value cut-off 0.05. Data are available in Supplementary Material 3. A cluster dendrogram of PQS characteristics was constructed in the program R, version 3.6.3, using the pvclust package to further reveal and graphically depict similarities between particular viral families. The following values were used as input data: Mean f (mean of predicted PQS per 1,000 nt), Min f (the lowest frequency of predicted PQS per 1,000 nt), Max f (the highest frequency of predicted PQS per 1,000 nt) and Cov % (% of the genome covered by PQS with a threshold of 1.2 or more) (Supplementary Material 4). The following parameters were used for both analyses: cluster method 'ward.D2', distance 'euclidean', number of bootstrap resampling was 10,000. Statistically significant clusters (based on approximately unbiased values above 95, equivalent to p-values less than 0.05) are highlighted by rectangles marked with broken red lines. R code is provided in Supplementary Material 4. Receiver operating characteristic (ROC) curves were constructed to determine the area under the curve (AUC) for genome groups and viral families causing either persistent or acute type of infection through web-based ROC [26]. Correlation was assessed by non-parametric Spearman's correlation coefficient. Normality of datasets was tested by d'Agostino-Pearson test.

2.4. Analysis of PQS around annotated NCBI features

Accessible features tables of all viral genomes were downloaded from the NCBI database. Features were grouped by their name stated in the feature table file. We analyzed PQS occurrence inside and around (before and after) predefined featured neighborhood (±100 nt). From this analysis, we obtained a file with feature names and number of PQS found inside and around features. Further processing was performed in Microsoft Excel and the data are available as Supplementary Material 5.

3. Results

3.1. Variation in PQS frequency in viruses and subviral agents

According to NCBI taxonomy classification, the analyzed genomes were divided into three main categories, viruses (10,613 genomes), satellites (321 genomes) and viroids (66 genomes) (Supplementary Material 6). Viruses were further divided according to their nucleic acid type into 7 groups (3,787 genomes of dsDNA viruses; 1,736 ssDNA; 438 dsRNA; 2,538 ssRNA+ viruses;

631 ssRNA-viruses; 132 ssRNA+ viruses with RT; 131 dsDNA viruses with RT) [27,28]. These groups were divided into 56 families as shown in the phylogenetic tree (Supplementary Material 6). Satellites were categorized into ssDNA (272 genomes) and ssRNA satellites (27 genomes). For further statistical analyses, only categories with 20 or more sequenced genomes were evaluated. Detailed phylogenetic classification for every analyzed virus and corresponding host is presented in Supplementary Material 1.

We analyzed the presence of PQS by G4Hunter in a pool of 13,809 sequences belonging to 11,000 unique viruses. All accessible genotypes and segments of each virus were considered. The detailed overall data are available in Supplementary Material 2. The smallest accessible complete genome was *Rice yellow mottle virus satellite* with the length of 220 nt, the longest one *Pandoravirus salinus*, a giant virus belonging to dsDNA viruses with a genome size of 2,473,870 bp [29]. The statistical review of results is shown in Table 1. The average GC content in all analyzed sequences was 45.29%; however, the GC content varied remarkably among the groups and families (Table 1) with a maximum of 57.98% GCs in *Pospiviroidae* (viroids) and a minimum of 33.72% GCs in *Poxviridae* (dsDNA viruses). For the whole pool of viruses, the mean PQS frequency with the default parameters of the analyses (window of 25 nt and score threshold of 1.2) was 1.13 per 1,000 nt and PQS cover on average 2.8% of the genomes. In comparison, the analysis of the human genome with the same parameters shows a PQS frequency of 1.71 with an average PQS coverage of 4.8%. The lowest frequency among the group category was found for ssRNA-viruses (0.67), followed by ssDNA satellites (0.96) and dsDNA viruses with RT (1.06). The highest frequency among the groups was observed for ssRNA+ viruses with RT (3.55) and ssRNA satellites (2.19). At the family level, the lowest mean PQS frequency was found in *Coronaviridae* (0.10) belonging to ssRNA+ viruses and *Poxviridae* (0.22) belonging to the dsDNA viruses and the highest for plant-infecting ssRNA+ viruses from the *Tymoviridae* (8.69) family, *Anelloviridae* (3.57) belonging to the ssDNA viruses and the *Retroviridae* family of ssRNA viruses with RT, where HIV-1 belongs (3.55). Detailed statistical analyses of inter group and inter family comparisons are provided in Supplementary Material 3.

Detailed statistical characteristics for PQS frequencies per 1,000 nt (including mean, variance, outliers) are depicted in boxplots for all inspected viral families (Fig. 1). The highest heterogeneity was found within the ssRNA+ viruses, where frequencies differed remarkably from 0.1 for the family of *Coronaviridae*; a family of important human pathogens to 8.69 inside the *Tymoviridae* family (*i.e.*, a 87-fold difference in PQS density). Within the ssDNA group, the highest PQS frequency was found for *Anelloviridae* (3.57), the family chronically infecting humans without known link to any disease [28] and the lowest for *Microviridae* (0.24), a family of bacteria-infecting viruses. A higher PQS frequency was observed for ssRNA+ viruses compared to ssRNA-ones. Two subgroups of dsDNA viruses with RT differed significantly: for *Caulimoviridae,* the plant infecting family of dsDNA viruses with RT, we observed a mean frequency of 0.44 PQS per 1,000 nt, while it was 1.75 for *Hepadnaviridae,* dsDNA viruses with RT to which hepatitis B virus (HBV) belongs.

To further demonstrate and graphically depict similarities among individual viral families we created a cluster dendrogram (Fig. 2). Input data and R code are provided in Supplementary Material 4. Interestingly, the families did not cluster according to their type of nucleic acid, as could be expected, but rather according to other phenotypic characteristics such as the dominant type of infection in Metazoa host. The significance of this clustering was confirmed by ROC analyses; while AUC values found for families causing either persistent or acute type of infection were set to 0.93 and 0.87 respectively, for genome groups they differed from 0.50 to 0.63 (Supplementary Material 4). Two evolutionary distinct families that typically cause persistent types of infection in *Metazoa* hosts, *Herpesviridae* (dsDNA) and *Retroviridae* (ssRNA with RT) [30–32], clustered together on the right side of the dendrogram with statistically significant similarity, whereas *Poxviridae* (dsDNA) clustered on the left side with another acutely infecting family *Coronaviridae* (ssRNA+) [33,34].

The virus with the highest PQS frequency (23.8 per 1,000 nt which translates into one PQS every 42 nt) was the *Grapevine fleck virus* with a GC content of 66.23%. This virus belongs to the *Maculavirus* genus in the order *Tymovirales*, in the family *Tymoviridae*. The RNA *Grapevine fleck virus* genome is linear, 7,564-nt long, and replicated by a positive stranded RNA virus replication model. Strikingly, PQS cover 66.17% of its genome when using a default threshold of 1.2 and 50.91% using a more stringent threshold of 1.4 (Fig. 3). In other words, the majority of the genome of this virus may potentially adopt a G4 fold. All PQS found in *Grapevine fleck virus* are located on the G-rich negative strand. ssRNA viruses are often non-symmetrically distributed with guanines and cytosines; in this case, guanines represent almost 50% of all bases on the negative intermediate. Moreover, PQS found in the genome of *Grapevine fleck virus* are remarkably long, with one PQS identified with a scanning window of 1,000 nt! Interestingly, this extreme density does not simply result from a high GC content, as viruses with an even higher proportion of guanines and cytosines, such as *Papiine alphaherpesvirus* (76.09%) have a lower PQS coverage (32.84%). This example again illustrates that the GC content is not strongly correlated with G4 propensity, as shown in other organisms such as Archaea [35]: in the *Grapevine fleck virus* this extremely unusual density in PQS is also the result a strong GC-skewness. Spearman's correlation coefficient for PQS frequency and GC content for viral sequences showed only a moderate positive correlation with a value of 0.65 (p-value < 0.001). After the G4Hunter analyses, the results were clustered according to the following intervals of G4Hunter score: 1.2–1.4, 1.4–1.6, 1.6–1.8, 1.8–2.0 and above 2.0 (Supplementary Material 7). PQS with higher G4Hunter score exhibit a higher probability of G4 formation [18,19]. The graphic representation of the GC contents and PQS frequencies is shown in Fig. 4. Presented PQS frequencies are normalized by virus with the highest PQS frequency in the corresponding G4Hunter score interval. All viruses with a high abundance of PQS relative to their GC content (over 70% of the maximal observed PQS frequency) are reported in Fig. 4. Different outliers could be seen depending on the analyzed intervals in Fig. 4: *Grapevine fleck virus*, *Cacao yellow mosaic virus* and *Grapevine asteroid mosaic associated virus, Grapevine red globe virus* and *Okra mosaic virus* which belong to *Tymoviridae* family of ssRNA+ viruses. The other outliers with more than 70% PQS belong to other classes of viruses: *Citrus exocortis viroid* and *Grapevine yellow speckle viroid 1* and *2* (viroids), *Human gyrovirus, Avian gyrovirus*, *Gyrovirus Tu789 and Micro torque teno virus* (ssDNA), *Equine encephalosis* (dsRNA) and *Ambe virus* (ssRNA-).

3.2. Localization of PQS in genomes of viruses

To evaluate the localization of PQS in viral genomes, we downloaded the genome annotations of all publicly accessible viral genomes and analyzed the presence of all PQS in each annotated *feature* and in its close proximity (100 nt before and after feature annotation). The pool of viral sequences was divided into 7 groups according to the type of nucleic acid as shown in Supplementary Material 6. Detailed results are summarized in Supplementary Material 5. The best annotated features in viral genomes are coding sequence (*CDS)* and *gene* (more than 371,000 annotated genes in viral genomes). The mean PQS frequency in *gene* was 0.9 per 1,000 nt, PQS frequency found in region 100 nt before *gene* was 1.02

Table 1
Genomic length, PQS frequencies and total counts. Seq (total number of sequences), Median (median length of sequences), GC % (average GC content), PQS (total number of predicted PQS), Mean f (mean of predicted PQS per 1,000 nt), Min f (the lowest frequency of predicted PQS per 1,000 nt), Max f (the highest frequency of predicted PQS per 1,000 nt) and Cov % (% of genome covered by PQS). Colors correspond to phylogenetic tree shown in Supplementary Material 6.

Domain	Seq	Median	GC%	PQS	Mean	Min	Max	Cov%
All	13,809	7,149	45.29	342,830	1.13	0	23.80	2.84
Viruses	13,405	7,328	45.33	342,130	1.13	0	23.80	2.83
Satellites	334	1,346	41.79	356	1.08	0	12.72	2.81
Viroids	66	339.5	56.68	49	2.09	0	16.26	2.83
Group	**Seq**	**Median**	**GC%**	**PQS**	**Mean**	**Min**	**Max**	**Cov%**
dsDNA	4,081	43,734	46.80	285,263	1.15	0	11.51	2.80
ssDNA	2,006	2,711	44.91	7,413	1.27	0	10.04	3.28
dsRNA	1,569	2,157	44.97	4,790	1.09	0	13.08	2.66
ssRNA+	2,818	7,511	45.97	25,570	1.21	0	23.80	3.10
ssRNA-	1,265	4,568	40.52	6,385	0.67	0	8.92	1.79
ssRNA+ RT	136	8,399.5	47.93	3,699	3.55	0	10.84	8.75
dsDNA RT	172	6,991	44.45	787	1.06	0	5.05	2.72
ssDNA satellites	285	1,350	40.21	296	0.96	0	7.43	2.50
ssRNA satellites	27	457	53.31	34	2.19	0	12.72	5.86
Family	**Seq**	**Median**	**GC%**	**PQS**	**Mean**	**Min**	**Max**	**Cov%**
Ackermannviridae	21	157,498	46.46	1589	0.48	0.35	0.72	1.23
Adenoviridae	189	35,172	52.55	21,945	3.27	0.16	7.57	7.91
Baculoviridae	91	131,160	41.74	2,903	0.25	0	1.81	0.57
Herpesviridae	120	153,119	53.14	66,125	3.43	0	11.44	8.91
Iridoviridae	40	114,073	46.49	8,323	1.82	0.08	3.62	4.31
Myoviridae	677	147,303	42.98	31,959	0.43	0	2.99	1.06
Papillomaviridae	369	7,673	43.34	5,618	1.97	0	6.04	5.00
Phycodnaviridae	36	192,531	41.35	6,031	0.69	0	1.88	1.69
Podoviridae	431	42,492	47.71	12,318	0.60	0	3.88	1.45
Polyomaviridae	142	5,088	42.04	1,277	1.68	0	7.81	4.63
Poxviridae	63	192,353	33.72	2,210	0.22	0	2.71	0.51
Siphoviridae	1,419	47,342	51.96	93,824	1.23	0	7.02	2.88
Anelloviridae	124	2,909	48.03	1,364	3.57	0.34	9.27	9.09
Circoviridae	245	1,901	46.17	891	1.87	0	10.04	4.76
Geminiviridae	894	2,737	43.72	1,999	0.83	0	4.72	2.18
Genomoviridae	117	2,199	51.54	601	2.33	0	6.71	6.15
Inoviridae	56	6,988	47.45	467	1.18	0	5.17	2.83
Microviridae	102	5,386	43.26	129	0.24	0	2.16	0.57
Parvoviridae	181	5,053	44.93	1,205	1.32	0	5.11	3.37
Smacoviridae	57	2,575	47.49	199	1.36	0	3.22	3.44
Amalgaviridae	21	3,412	48.75	137	1.91	0	5.84	5.07
Endornaviridae	38	14,157.5	41.54	199	0.34	0	1.08	0.85
Partitiviridae	162	1,832	46.48	569	1.79	0	11.19	4.38
Reoviridae	1,060	1,876.5	42.95	1,737	0.84	0	13.08	2.06
Totiviridae	99	5,193	51.51	1,034	1.86	0	7.73	4.61
Alphaflexiviridae	61	6,656	50.08	991	2.45	0	10.93	5.92
Arteriviridae	24	15,235.5	51.70	433	1.20	0.45	2.72	3.10
Astroviridae	70	6,613.5	47.19	421	0.93	0	5.21	2.26
Betaflexiviridae	120	8,390	43.90	658	0.70	0	2.88	1.67

Bromoviridae	134	2,862	43.95	128	0.36	0	1.89	0.94
Caliciviridae	399	7,511	50.41	5,662	1.89	0.35	5.06	5.09
Coronaviridae	208	29,695	39.78	575	0.10	0	0.52	0.23
Closteroviridae	72	13,227	40.99	480	0.51	0	2.28	1.33
Dicistroviridae	27	9,275	39.05	67	0.27	0	0.86	0.71
Flaviviridae	221	10,560	51.58	4,446	2.04	0.26	5.63	5.03
Iflaviridae	45	9,932	38.87	204	0.46	0	1.56	1.27
Luteoviridae	55	5,808	49.07	527	1.68	0	3.23	4.66
Narnaviridae	63	2,718	41.22	105	0.56	0	4.37	1.34
Picornaviridae	435	7,413	45.04	4,282	1.25	0	10.70	3.34
Potyviridae	185	9,689	42.51	452	0.28	0	1.75	0.70
Secoviridae	138	5,841.5	44.10	498	0.63	0	2.72	1.72
Solemoviridae	20	4,172.5	50.09	93	1.09	0	2.65	2.67
Togaviridae	44	11,620	51.23	452	0.91	0.25	6.15	2.25
Tombusviridae	84	4,016	49.54	601	1.93	0	7.27	4.81
Tymoviridae	46	6,334	55.50	2363	8.69	1.63	23.80	21.86
Virgaviridae	85	6,311	42.39	165	0.38	0	1.96	0.88
Arenaviridae	96	3,534	42.09	450	0.96	0	3.20	2.32
Hantaviridae	105	6,637	38.54	174	0.49	0	2.17	1.39
Orthomyxoviridae	184	1,721	43.49	214	0.68	0	4.05	1.65
Paramyxoviridae	92	15,621	42.91	1,754	1.20	0.13	3.71	2.97
Peribunyaviridae	197	4,527	36.75	182	0.37	0	5.08	1.09
Phenuiviridae	144	3,402.5	41.19	302	0.60	0	4.39	2.03
Rhabdoviridae	184	11,902.5	41.66	2,123	0.98	0	3.32	2.53
Retroviridae	136	8,399.5	47.93	3,699	3.55	0	10.84	8.75
Caulimoviridae	91	7,586	41.05	309	0.44	0	1.31	1.09
Hepadnaviridae	81	3,215	48.27	478	1.75	0	5.05	4.55
Alphasatellitidae	102	1,361	41.59	105	0.80	0	3.38	1.96
Tolecusatellitidae	148	1,351.5	39.18	129	0.72	0	7.43	1.97
Pospiviroidae	52	329.5	57.98	36	1.99	0	16.26	4.45

and 100 nt after *gene* 1.04 PQS per 1.000 nt. The best annotated in viral genomes is the "gene" feature. To compare localization in various features we took PQS frequency in genes as "1" and compare enrichment or depletion in other locations relative to this normalized value. PQS frequency in every feature type was calculated first for each virus separately and then the average was made. PQS frequencies in comparison to PQS frequencies in *gene* in annotated genome sites and in their close proximity are shown in Fig. 5. Statistically significant differences between PQS frequency in gene and other features in the corresponding category are highlighted by asterisk. In the first group, dsDNA viruses, the most notable enrichment was noted inside and around *repeat regions* (5.1- and 1.6-times more frequent compared to gene). A higher PQS frequency in comparison to *gene* average PQS abundance was observed also inside and after *misc_RNAs* (2.9- and 3.5-times, respectively), *tmRNA* (2-times) and *exons* (2.2-times more frequent than gene). In ssDNA viruses, similarly to dsDNA group, the highest PQS frequency was found inside *repeat regions* (9.2-times higher compared to *gene* region). In contrary to dsDNA group, but in agreement with dsRNA viruses, an abundance of PQS was observed also in *stem_loops* (6.5- and 2.1-times more) and *5′UTRs* (2.5- and 1.6-times). Higher frequency in comparison to PQS frequency in *genes* was found also inside the *introns* (2.5-times higher) and *rep_origins* (1.8-times). For ssRNA+ group, the most significant enhancement was found around *exon* (1.7- and 3.4-times, respectively), before *regulatory elements* (1.9) and inside *miscRNAs* (1.6). For ssRNA-group, no PQS were found inside *5′UTRs*, but the enrichment was noted also inside *stem_loops* (35.3-times), *misc_RNA* (2.1-times), *regulatory elements* (2.2-times) and before and in *exons* (4.4- and 2.3-times). For ssRNA viruses with RT, the highest PQS frequency compared to *gene* was observed before *exons* (3.3-times), however, the abundance of PQS was observed also inside *misc_binding* sites (2.3-times) *3′UTRs* and *5′UTRs* (1.4-times). In dsDNA viruses with RT the most notable enrichment was before *mRNAs* (4-times) after *regulatory elements (1.8-times respectively) and* around *misc_RNAs (2.1- and 3.3-times)*. Satellites are enriched for PQS in *stem_loops* (14.9-times) and *repeats* (22.7-times).

3.3. Comparison of PQS frequency in viruses with dominant acute and persistent type of the infection

The cluster and ROC analyses (Fig. 2) revealed that the type of nucleic acid is not the main factor influencing the PQS prevalence, as might be expected, but viruses rather cluster according to other phenotypic characteristics, such as dominant type of infection [31,32,36,37]. In general, a viral infection is considered acute when onset and resolution of infection is rapid, or persistent when primary infection is not cleared by the immune response and virus

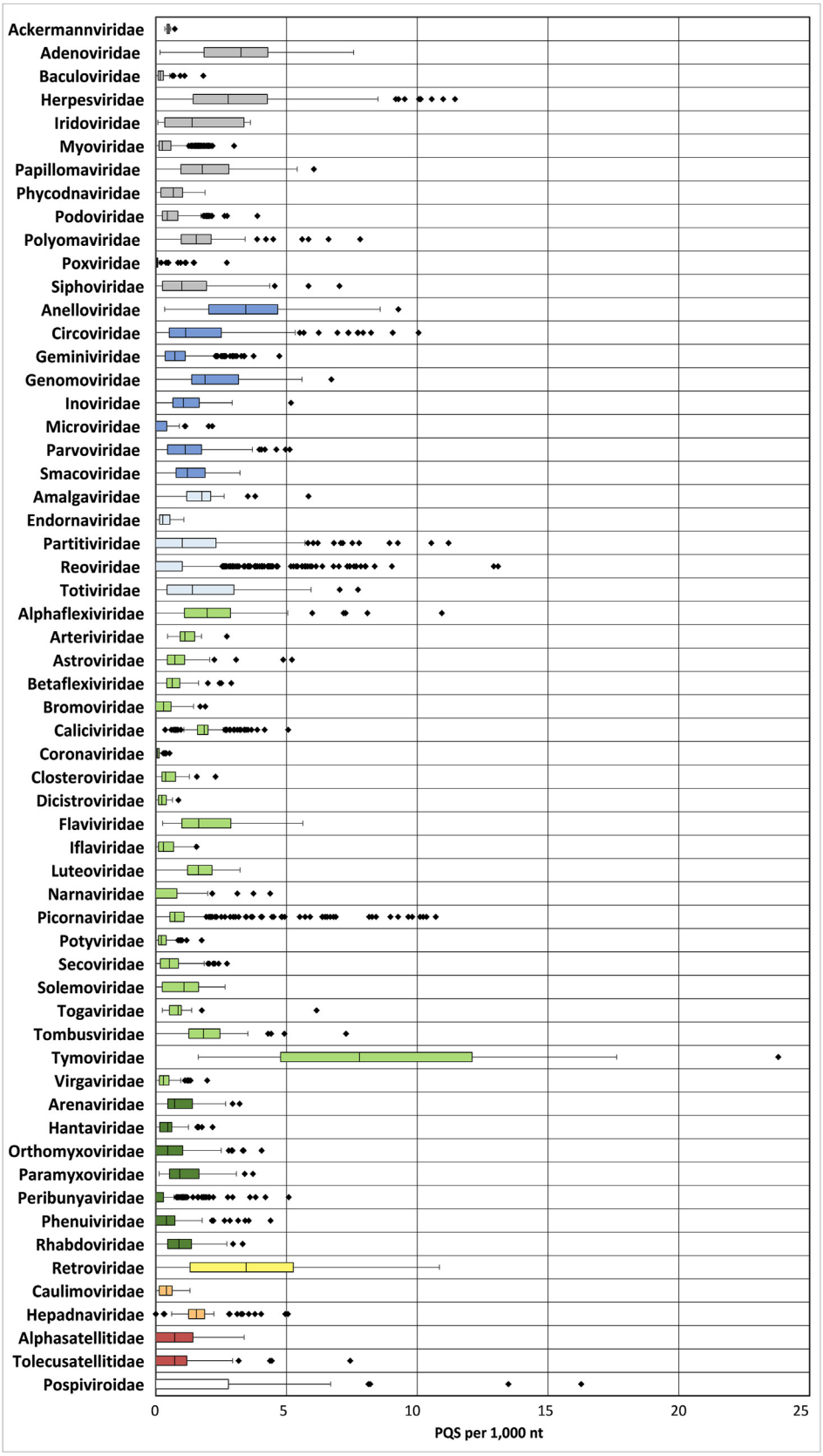

Fig. 1. Frequencies of PQS in families of the analyzed viral genomes. Data within boxes span the interquartile range and whiskers show the lowest and highest values within 1.5 interquartile range. Black diamonds denote outliers. Colors correspond to the phylogenetic tree shown in Supplementary Material 6.

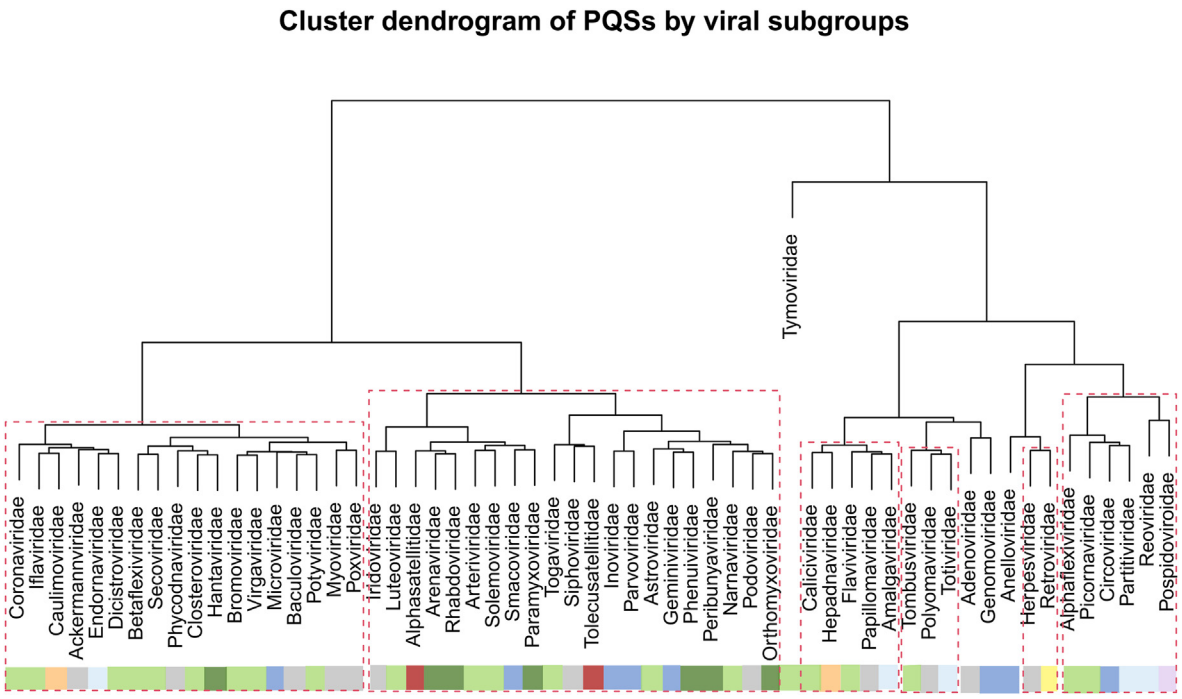

Fig. 2. Cluster dendrogram based on PQS characteristics in all viral families. Input data are listed in Supplementary Material 4. Statistically significant clusters (based on AU values above 95, equivalent to p-values lower than 0.05) are highlighted by rectangles drawn with dashed lines. Colors corresponds to phylogenetic tree in Supplementary Material 6.

remains in host cells either as a provirus or episome. This distinction is not always straightforward, as some viruses causing acute infection may also establish persistency in immunocompromised hosts [38]. Based on the literature [30–34,36–49] we chose only extreme groups of viruses which typically cause either persistent or acute type of infection in *Metazoa* hosts and separated them in two main groups: 1. those causing persistent infections (*Adenoviridae, Herpesviridae, Papillomaviridae, Polyomaviridae, Anelloviridae, Hepacivirus* genus of *Flaviviridae* family, *Retroviridae*, and *Hepadnaviridae*); 2. viruses causing mainly acute infections (*Poxviridae, Rotavirus* genus of *Reoviridae*, *Coronaviridae*, *Hepatovirus* genus and Rhinoviruses of *Picornaviridae* family and *Orthomyxoviridae*). In case the family of viruses was too diverse, we selected lower taxonomic units. For example in *Flaviviridae* family; while viruses from genus *Hepacivirus* are a common cause of chronic infections, the genus *Flavivirus* contain viruses with a much more diverse behavior [50,51]. Similarly *Reoviridae* family includes the acutely infecting genus *Rotavirus,* but also other genera of plant pathogens [52]. For further statistical analyses, only groups with 20 or more sequenced genomes were evaluated. PQS frequencies are shown in Fig. 6, panel A. While the persistently infecting viruses have an average PQS frequency of 2.66 PQS per 1,000 nt with five groups with a PQS frequency above 3 PQS per 1000 nt (*Adenoviridae* (3.27), *Herpesviridae* (3.55), *Anelloviridae* (3.57), *Hepacivirus* (3.31), *Retroviridae* (3.55), all average PQS frequencies of acutely infecting groups of viruses are below 0.7 per 1,000 nt (average of 0.33, with values ranging from 0.04 for *Rotaviruses* and 0.68 for *Orthomyxoviridae*; *Poxviridae* (0.22), *Coronaviridae* (0.097), *Hepatoviruses* (0.6), and Rhinoviruses (0.57) falling in between). The values found for the viruses in the second group are statistically significantly different from the density of group 1 viruses, independently of the type of nucleic acid (Fig. 6, panel B and Supplementary Material 3). PQS frequencies therefore correlate with the dominant type of infection rather than with the type of nucleic acid (dsDNA virus from *Adenoviridae, Herpesviridae* vs dsDNA viruses of *Poxviridae* as shown in Fig. 2), intracellular localization (nucleus localization of persistently infecting *Herpesviruses* and acutely infecting *Orthomyxoviridae* vs cytoplasmatic localization of acutely infecting *Poxviridae* and chronically infecting *Hepacivirus*) or tissue tropism. Such example is viruses causing hepatitis; whereas hepatitis A virus (HAV) from *Hepatovirus* of *Picornaviridae* family with ssRNA+ genomes cause only acute hepatitis with average PQS frequency of 0.6, HBV from *Hepadnavirus* of *Hepadnaviridae* (dsDNA RT) and hepatitis C virus (HCV) from *Hepacivirus* of *Flaviriridae* (ssRNA+) are causative agents of chronic hepatitis with average PQS frequency of 1.75 and 3.31, respectively (Fig. 6, panel C). Moreover, both HAV and HCV are ssRNA+ viruses with statistically significant difference in PQS frequency. Hepatitis D virus (HDV) and hepatitis E virus (HEV) were not included in the division as HDV is a satellite of HBV as it requires the HBV envelope proteins to form virus particles, and the infection type of HEV is not straightforward [53–55].

4. Discussion

The presence of the G4s in viral genomes has attracted eminent interest due to their potential in antiviral therapies [6]. The most comprehensive viral genome analyses to date of PQS have been done using a pattern-based algorithm [23,24]. However, pattern-based algorithms do not consider atypical G4 forming structures, including bulges for example [56]. In this study, we have broadened the range of analyses significantly by looking at *all* sequenced viral genomes, not only human-hosted viruses (among 11,000 viruses studied, 1,184 are human hosted). The presence and locations of PQS in more than thirteen thousand viral sequences were analyzed using G4Hunter, which allows the identification of non-canonical PQS and the categorization of results based on G4Hunter score values. By expanding the host spectrum to all organisms, we gained crucial additional information about possible zoonotic viral pathogens, conservation of PQS among different species and

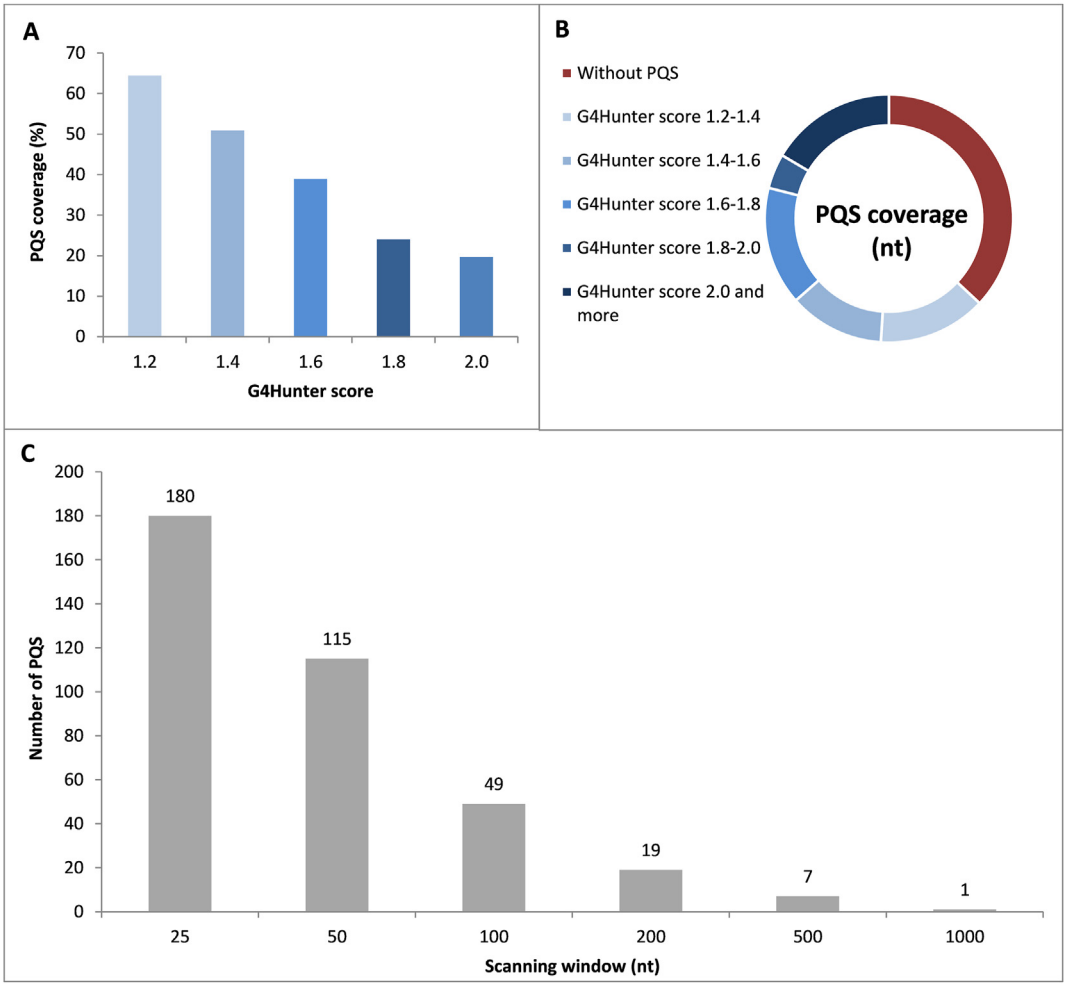

Fig. 3. Graphical visualization of PQS densities in the *Grapevine fleck virus*. (A) PQS coverage of the *Grapevine fleck virus* genome with the different G4Hunter score threshold. (B) Representation of the PQS length in the *Grapevine fleck virus* according to different G4hunter score intervals' (C) Number of PQS found according to different scanning window length using a threshold of 1.2.

interspecies variability in PQS frequency. Moreover, we broadened the analyses also to subviral agents. On average, viroids showed higher PQS frequency (2.09) than viruses (1.13) and satellites (1.08). Since viroids do not code for any protein, it has been proposed that their RNA secondary structures may play important roles in their life cycle [20].

Grapevine fleck virus, *Cacao yellow mosaic virus, Grapevine red globe virus* and *Okra mosaic virus* are four outliers found in the GC content analyses (Fig. 4). All of them belong to the *Tymoviridae* family of ssRNA+ viruses, the family with the highest PQS frequency found (8.69). *Grapevine fleck virus* has the highest frequency of 23.8 PQS per 1,000 nt among all analyzed genomes. Such frequency is extraordinary, especially for ssRNA+ viruses, where the average PQS frequency is only 1.21 PQS per 1,000 nt. The *Grapevine fleck virus* is strictly confined to the phloem of infected hosts. Infected cells contain multivesiculate bodies, cytopathic structures derived from mitochondria, and a great quantity of accumulated viral particles, sometimes also in crystalline arrays [57]. The *Anelloviridae* family has the second highest PQS frequency (3.57) among all families (Table 1, Fig. 1). Viruses of the *Anelloviridae* family cause chronic human infections without any known direct clinical symptoms [30]. Four viruses of this family, *Human gyrovirus*, *Avian gyrovirus, Gyrovirus Tu789* and *Micro Torque teno virus,* exhibited extremely high PQS frequencies (Fig. 4). The third highest PQS frequency among genome subgroups was found for the *Retroviridae* family (3.55), which includes RNA viruses with reverse transcriptase, as shown in Table 1. G4s in genomes of the *Retroviridae* family, especially in HIV-1 virus from the genus *Lentivirus*, have been intensively studied with respect to promising antiviral therapy via G4s targeting. Three potential G4-forming sequences, LTR-II, LTR-III and LTR-IV, are localized in highly conserved sp1 and NFκB binding sites of the U3 region of the long terminal repeat (LTR). Viral transcription can be regulated by G4 stabilizing proteins and ligands. Whereas selective stabilization of LTR-III leads to inhibition of viral transcription, stabilization of LTR-IV has the opposite effect [58,59]. Our analyses confirm that PQS of *Retroviridae* family are enriched for PQS in *3′UTRs and 5′UTRs* compared to *genes*, as shown in Fig. 5. PQS was found also in the *nef* gene of HIV virus, an important virulence factor, responsible for T-cell activation and maintenance of a persistent state of infection [60].

Among dsDNA viruses, the highest observed PQS frequency was found inside the *Herpesviridae* family (3.43). The importance of the roles of G4s in the regulation of basic molecular mechanisms was reported for several species of this family [11,61–64]. G4 stabilizing compounds could inhibit replication, gene expression and importantly alter viral life cycle [5,65]. Our analyses indicate that PQS are mostly found in *repeat regions, misc_RNA* and *tmRNA*s. Another dsDNA family with significantly higher PQS frequency is

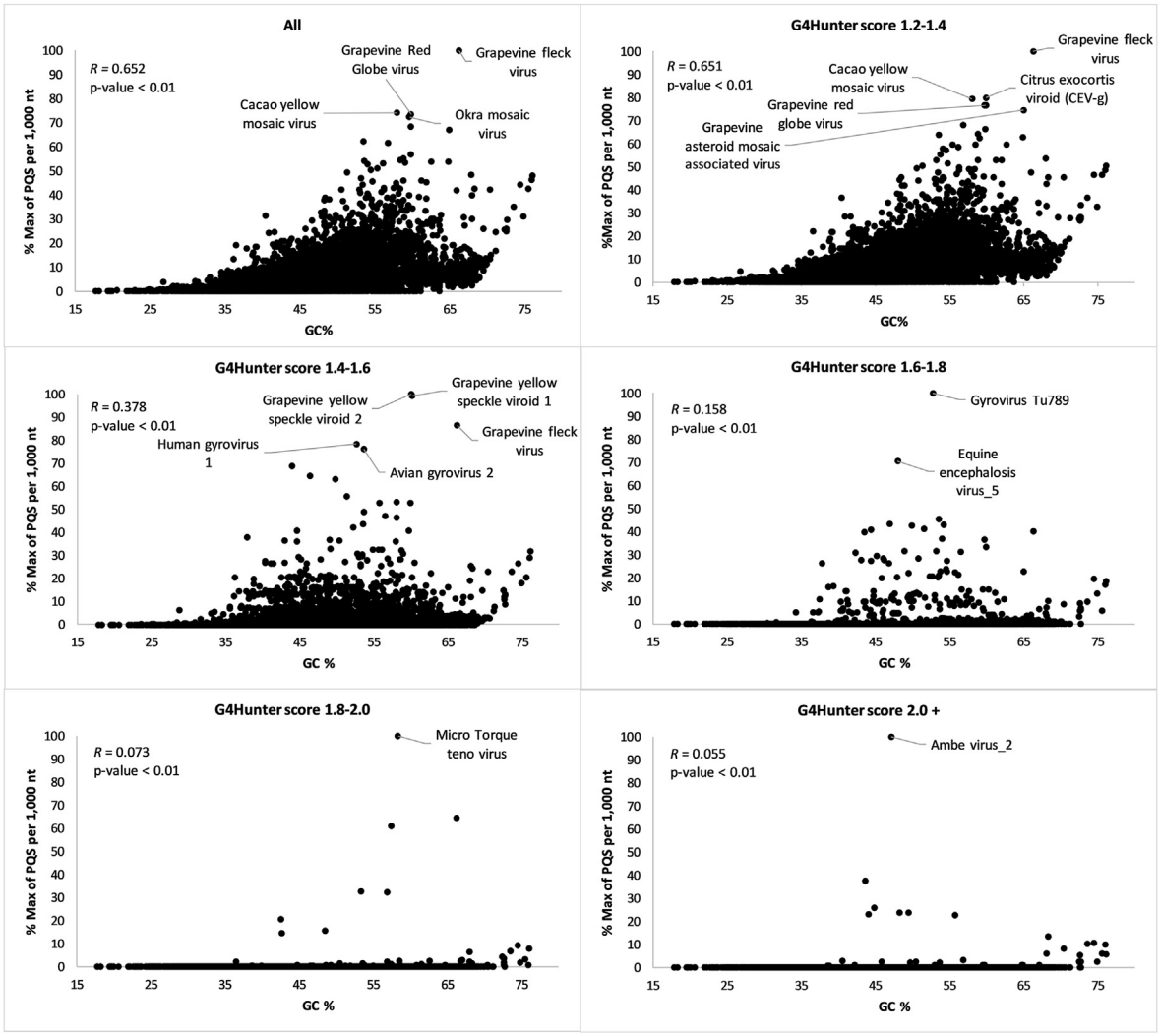

Fig. 4. Relationship between observed frequency of PQS per 1,000 nt and GC content in all analyzed viral sequences in various G4 Hunter score intervals. In each G4Hunter score interval miniplot, frequencies were normalized according to the highest observed frequency of PQS. The names of viruses exhibiting a relative maximal frequency per 1,000 nt greater than 70% are reported. Information about statistical evaluation by Spearman's correlation analyses are shown in the graphs.

Adenoviridae. Genome-wide analyses revealed 15 G4s to be conserved across different species of human adenoviruses [66]. Moreover, the G4 ligand BRACO-19 inhibited multiplication of the virus in human cells [66]. For both families, *Herpesviridae* and *Adenoviridae*, the presence of G4s was found in early-expressed genes, which encode regulatory proteins crucial for the next phases of viral life cycles [66,67]. Inside the *Papillomaviridae* family (where we noted the third highest PQS frequency (1.97) among dsDNA viruses), G4 forming sequences were found by pattern based algorithm in 8 out of the 120 identified HPV genotypes, including oncogenic high risk *Human Papillomavirus type 52* (HPV-52), HPV-58 [68]. Our analyses identified only 5 human genotypes (out of 188) completely lacking PQS. It has recently been shown that PQS frequency in dsDNA viruses correlate with PQS frequencies in corresponding hosts, pointing to evolutionarily reciprocal mimicking of virus-host genome organization [69].

A higher frequency of PQS was found for ssRNA+ viruses in comparison to ssRNA-viruses. For both groups, both positive and negative strands were analyzed for the presence of PQS as both strands are essential in the life cycle of the virus. For example, the stabilization of G4 formation on the negative intermediate of HCV (belonging to the *Flaviviridae* family of ssRNA+ viruses) can modulate RNA synthesis [70]. Localization of PQS was evaluated in all genome groups with emphasis on annotated features. Several groups of viruses exhibited a PQS abundance in *5′UTRs*, *3′UTRs*, *repeats*, or *replication origin* compared to gene locations (Fig. 5). Enrichment in these particular areas of the genomes suggest important roles of G4s in regulation of different basic molecular processes [6].

Based on our analyses we demonstrate that PQS frequency in viral genomes is correlated to the type of viral infection in *Metazoa* hosts. We observed statistically significant higher PQS frequencies (as shown in Fig. 6, Supplementary Material 3) in subgroups of viruses that have previously been reported [30–32,36,40–43,49] to cause persistent, chronic or latent infections in *Metazoa* hosts such as *Adenoviridae*, *Herpesviridae*, *Papillomaviridae*, *Polyomaviridae*, *Anelloviridae*, *Hepacivirus*, *Retroviridae* or *Hepadnaviridae*. On the contrary, low PQS frequencies were observed in subgroups causing mostly acute infections [33,34,36,37,45,48], such as *Poxviridae*, *Rotavirus* genus of *Reoviridae* family, *Hepatovirus* genus and Rhinoviruses of *Picornaviridae* family, the *Coronaviridae* family of ssRNA+ viruses, as well as the *Orthomyxoviridae* family of ssRNA- viruses. Interestingly, PQS frequency correlates with the type of infection rather than with the type of nucleic acid, phylogenetic

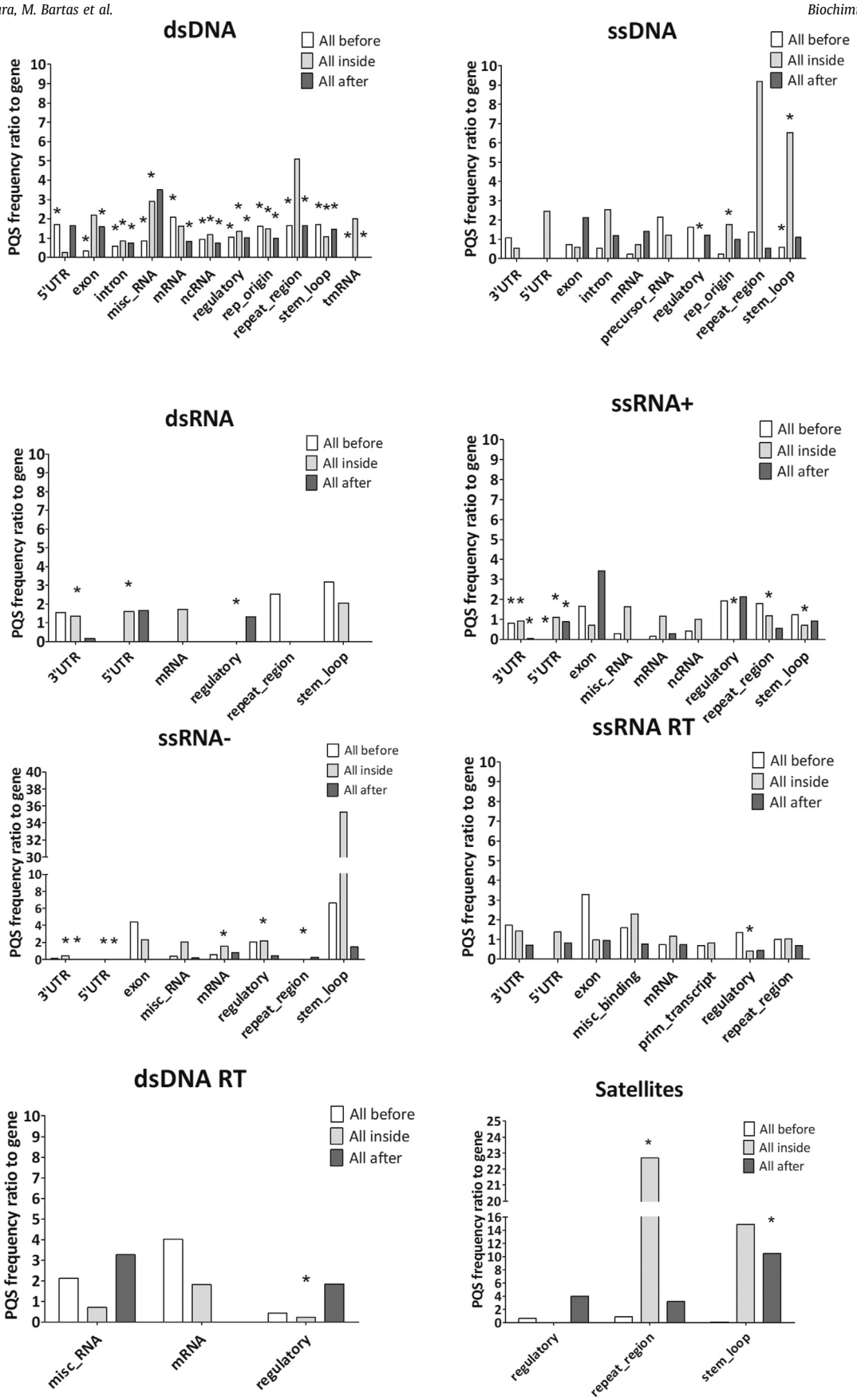

Fig. 5. Differences in PQS frequency by annotated locus. The chart shows the ratio of PQS frequencies per 1,000 nt between *gene* annotation and other annotated locations from the NCBI database. We analyzed the frequencies of all PQS within (inside), before (100 nt) and after (100 nt) annotated locations. Detailed results are summarized in Supplementary Material 5. Asterisks denote statistically significant differences (p-value < 0.05) between PQS frequency in *gene* and other feature in corresponding category evaluated by non-parametric Kruskal-Wallis test and the results are presented in Supplementary Material 3.

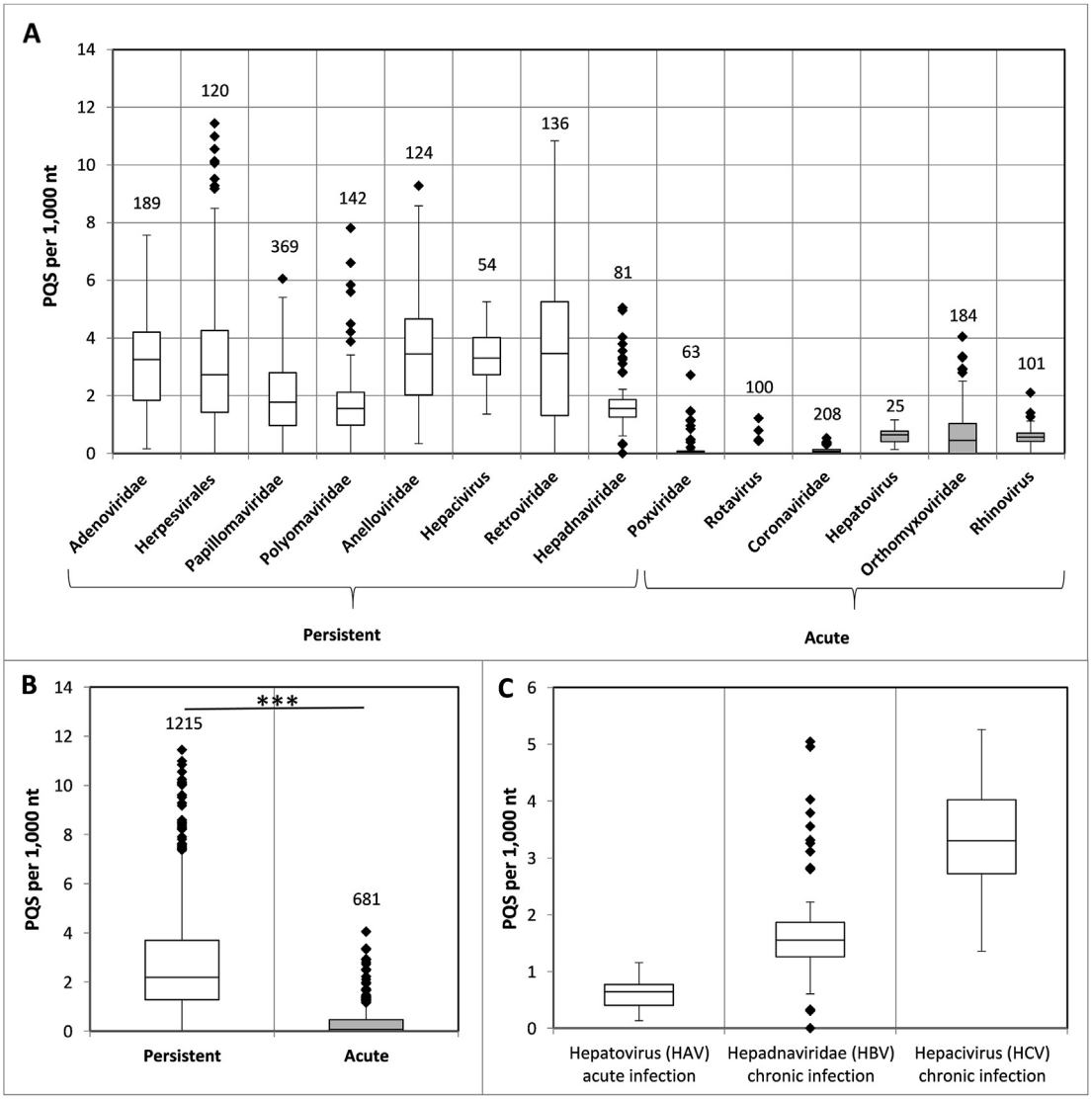

Fig. 6. Comparison of PQS frequency in viruses with dominant persistent and acute types of the infection. Data within boxes span the interquartile range and whiskers show the lowest and highest values within 1.5 interquartile range. Black diamonds denote outliers and the digits indicate the number of viruses in each category. Selected groups of viruses were categorized according dominant type of the infection (panel A). Viruses causing mostly persistent infection (white) are statistically significantly (p-value 3E-219) enriched in PQS compared to acutely infecting viruses (grey) (panel B). PQS frequencies of three viral groups causing either acute or chronic hepatitis (panel C). Statistical evaluation of the data set has been made through Kruskal-Wallis test and the results are presented in Supplementary Material 3.

categorization, or tissue tropism. For example, viruses belonging to the *Poxviridae* family with a dsDNA genome that cause acute infections have a significantly lower PQS frequency than other dsDNA subgroups that are known to cause persistent infections (such as *Adenoviridae* and *Papillomaviridae* family or order *Herpesviridae*) [31,34]. ssRNA+ viruses from genus *Hepacivirus* that cause chronic hepatitis and hepatocellular carcinoma in humans [66] have significantly higher PQS than ssRNA+ viruses from genus *Hepatovirus*, acutely infecting viruses with same tissue tropism [36]. Moreover, cluster analyses (Fig. 2) revealed that viruses causing persistent infections clustered together regardless of the nucleic acid type and current phylogenetic classification. On top of that, two evolutionary distinct subgroups that typically cause persistent types of infection, *Herpesviridae* and *Retroviridae*, clustered together with statistically significant similarity (Fig. 2). However, identifying the primary type of infection is not always straightforward, as found for the *Filoviridae* family to which Ebola and Marburg virus belong. Members of this family with an average PQS frequency of 0.95 PQS per 1,000 nt cause severe acute infections with extreme mortality rates over 50%. However, both Ebola and Marburg virus RNA have been found in semen, cerebrospinal and intraocular fluids of survivors long after blood samples tested negative [71,72]. Stabilization of G4 inhibits L gene expression, which encodes for RNA-directed RNA polymerase of Zaire ebolavirus [73,74]. Another example of not clear category of viruses could be Measles virus of *Paramyxoviridae* family with the 2.33 PQS per 1,000 nt causing mostly acute infection, but it can persist in the nervous system and was linked to chronic disease such as Paget's disease and otosclerosis [75]. The presence and high prevalence of G4 thus cannot be considered as the only indicator of the type of infection, but particular G4s could play a role in a general viral persistency in Metazoan hosts.

Evolutionary elimination of G4s may represent a selective advantage for actively replicating viruses, for at least two reasons: i) this would avoid any problem in replication, transcription or reverse transcription, as stable DNA and RNA G4s may interfere

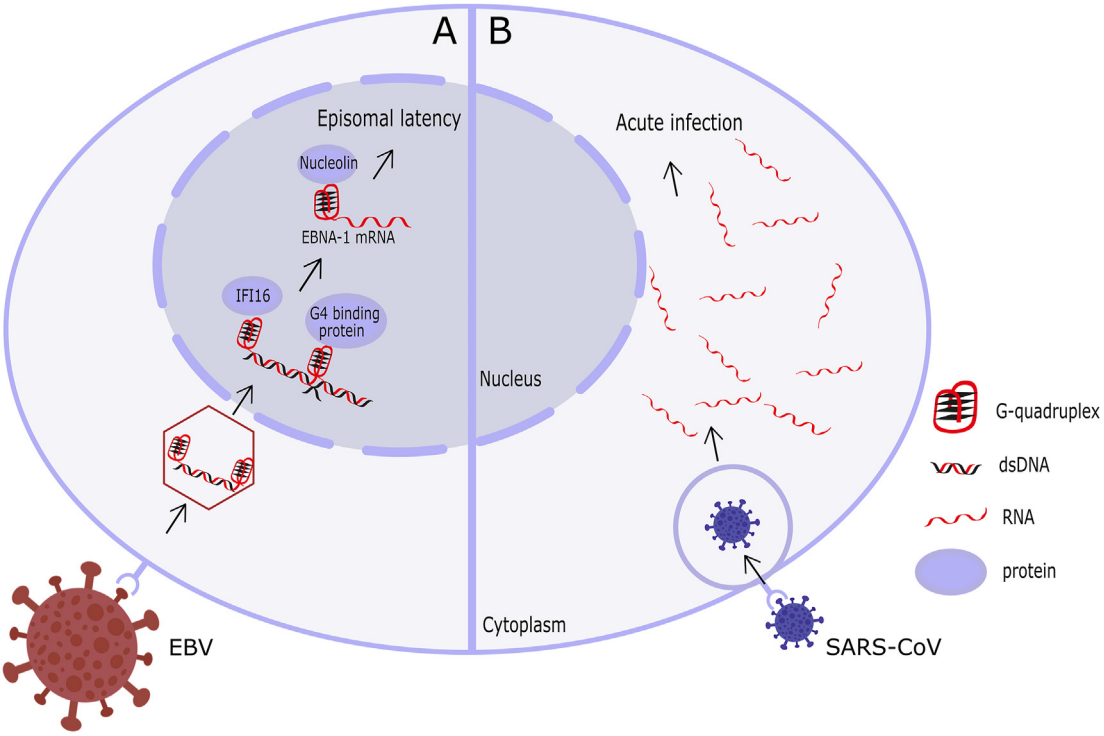

Fig. 7. Schematic model of G4 influence on the viral life cycle. Viruses that cause persistent infections, such as EBV, have high potential PQS frequency (panel A) that may be recognized by DNA sensors (e.g. IFI16) [76,77] and other G4-binding proteins (e.g. nucleolin) [11]. Binding to a G4-binding protein such as nucleolin can influence latency and immune evasion, as found for EBNA-1. In contrast, viruses that cause acute infections, such as SARS-CoV, tend to have a low PQS frequency (panel B).

with these processes; ii) their genome would not be recognized by cellular G4 binding proteins and alien DNA sensors of the host cells, such as interferon-γ-inducible protein 16 (IFI16) [11,76], a protein which preferentially binds to G4 and is critical for maintenance of EBV latency [11,76,77]. G4s depletion thus could be one of the reasons why these viruses are able to quickly develop an acute phase of infection. On the other hand, viruses with high PQS frequency would be forced to "hide" their genomes into the host genome, which leads to development of the persistent phase of the infection (Fig. 7). Acutely infecting viruses are the causes of seasonal epidemics. This hypothesis confirmed also contemporary analyses of H1N1 flu genome [78] and SARS-CoV-2 genome in which PQS are rare [79,80]. So far, the involvement of G4s in the regulation of viral life cycles has been demonstrated on HIV-1 virus from *Retroviridae* family, Kaposi's sarcoma-associated herpes virus (KSHV) and EBV virus belonging to *Herpesviridae* family. In the case of HIV-1 virus, treatment of latently infected cells with G4 ligands BRACO-19 and TMPyP4 led to inhibition of HIV-1 replication and increased apoptosis rate [81]. Moreover, formation of the three conserved G4s has been confirmed in the *nef* gene of HIV-1, an important factor in maintenance of the persistent state of infection. Treatment with a G4 ligand reduced *nef* expression and repressed nef-dependent HIV-1 infectivity [60]. G4 is also located inside the regulator of latency and immune evasion of another virus – herpesvirus EBV. EBV, formally known as a *Human gammaherpesvirus 4,* is a causative agent of different lymphoproliferative disease, such as Burkitt lymphoma or Hodgkin lymphoma. EBNA-1 is the only EBV protein required for replication and segregation of episomal genomes in latency, resulting in the maintenance of the EBV genomes. EBNA-1 is also highly antigenic; destabilization of the G4 structures through codon-modification significantly enhanced *in vivo* antigen presentation and activation of virus-specific T cells [82]. Formation of RNA G4 in EBNA1 mRNA of EBV inhibits its translation and, as a result, also the presentation of antigen. G4 formation in mRNA of EBNA1 therefore represents a mechanism of how the virus balances between latency, immune evasion and stable copy number of genomes [11,65]. The analogous gene in the KSHV genome, *LANA-1*, is the most abundantly expressed protein during latency which is required for the persistence of KSHV in the host cell. It has been demonstrated that stabilization of G4 in LANA-1 mRNA inhibits its translation and immune recognition [83]. KSHV virus, also known as *Human gammaherpesvirus 8* causes Kaposi sarcoma and other malignancies, mostly in immunodeficient patients. G4 ligands could eliminate even extrachromosomal episomes of KSHV [5]. Overall, while the presence of PQS has been associated with latency for some viruses, our study provides for the first time a statistical link between G4 propensity and persistence of the virus.

5. Conclusions

In this article, we analyzed the presence of PQS in all accessible viral genome sequences. The number of PQS differs remarkably among individual groups and families, suggesting evolutionary adaptations connected with G4. The most striking result was the observation that the PQS frequency varied according to the type of infection, rather than to nucleic acid type of the genome: while viruses typically causing persistent infections in *Metazoa* hosts require a similar G4 background as the host and are enriched for the presence of PQS, the viruses causing acute phase only are often reduced or absent in PQS. We have shown that PQS are non-randomly distributed in viral genomes and associated with specific regions in their genomes, which have mainly regulatory functions. These data allow use of the unique and non-random localization of PQS in viral genomes in their characterization and points to G4 importance for viral lifecycles.

Funding

This work was supported by The Czech Science Foundation (18-15548S), the ANR Flash-Covid 2021 call and by the SYMBIT project Reg. no. CZ.02.1.01/0.0/0.0/15_003/0000477 financed from the ERDF.

Author contributions

Conceptualization, V.B.; Data curation, N.B and A.C.; Formal analysis, N.B. and M.B.; Funding acquisition, V.B. and M.F.; Investigation, N.B and A.C.; Methodology, V.B.; Project administration, V.B.; Resources, P.K. and J.Š.; Software, P.K. and J.Š.; Supervision, V.B., M.F., and P.P.; Validation, V.B.; Visualization, N.B., A.C. J.L.M. and M.B.; Writing—original draft, N.B., A.C, M.B. and V.B.; and Writing—review and editing, P.P., J.L.M. and M.F.

Declaration of competing interest

The authors declare no competing interests.

Acknowledgements

We thank Prof. Richard Bowater, University of East Anglia, for proofreading and providing feedback on the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.biochi.2021.03.017.

References

[1] R.W. Harkness, A.K. Mittermaier, G-quadruplex dynamics, Biochim. Biophys. Acta Protein Proteonomics 1865 (2017) 1544–1554, https://doi.org/10.1016/j.bbapap.2017.06.012.
[2] D. Rhodes, H.J. Lipps, G-quadruplexes and their regulatory roles in biology, Nucleic Acids Res. 43 (2015) 8627–8637, https://doi.org/10.1093/nar/gkv862.
[3] Y. Ding, A.M. Fleming, C.J. Burrows, Case studies on potential G-quadruplex-forming sequences from the bacterial orders Deinococcales and Thermales derived from a survey of published genomes, Sci. Rep. 8 (2018) 1–11, https://doi.org/10.1038/s41598-018-33944-4.
[4] M. Falabella, J.E. Kolesar, C. Wallace, D. de Jesus, L. Sun, Y.V. Taguchi, C. Wang, T. Wang, I.M. Xiang, J.K. Alder, R. Maheshan, W. Horne, J. Turek-Herman, P.J. Pagano, C.M. St Croix, N. Sondheimer, L.A. Yatsunyk, F.B. Johnson, B.A. Kaufman, G-quadruplex dynamics contribute to regulation of mitochondrial gene expression, Sci. Rep. 9 (2019) 5605, https://doi.org/10.1038/s41598-019-41464-y.
[5] A. Madireddy, P. Purushothaman, C.P. Loosbroock, E.S. Robertson, C.L. Schildkraut, S.C. Verma, G-quadruplex-interacting compounds alter latent DNA replication and episomal persistence of KSHV, Nucleic Acids Res. 44 (2016) 3675–3694, https://doi.org/10.1093/nar/gkw038.
[6] E. Ruggiero, S.N. Richter, Viral G-quadruplexes: new frontiers in virus pathogenesis and antiviral therapy, Annu. Rep. Med. Chem. 54 (2020) 101–131, https://doi.org/10.1016/bs.armc.2020.04.001.
[7] S. Asamitsu, S. Obata, Z. Yu, T. Bando, H. Sugiyama, Recent progress of targeted G-quadruplex-preferred ligands toward cancer therapy, Molecules 24 (2019) 429, https://doi.org/10.3390/molecules24030429.
[8] S.K. Mishra, A. Tawani, A. Mishra, A. Kumar, G4IPDB: a database for G-quadruplex structure forming nucleic acid interacting proteins, Sci. Rep. 6 (2016) 38144, https://doi.org/10.1038/srep38144.
[9] V. González, L.H. Hurley, The C-terminus of nucleolin promotes the formation of the c-MYC G-quadruplex and inhibits c-MYC promoter activity, Biochemistry 49 (2010) 9706–9714, https://doi.org/10.1021/bi100509s.
[10] E. Tosoni, I. Frasson, M. Scalabrin, R. Perrone, E. Butovskaya, M. Nadai, G. Palù, D. Fabris, S.N. Richter, Nucleolin stabilizes G-quadruplex structures folded by the LTR promoter and silences HIV-1 viral transcription, Nucleic Acids Res. 43 (2015) 8884, https://doi.org/10.1093/nar/gkv897.
[11] M.J. Lista, R.P. Martins, O. Billant, M.-A. Contesse, S. Findakly, P. Pochard, C. Daskalogianni, C. Beauvineau, C. Guetta, C. Jamin, M.-P. Teulade-Fichou, R. Fåhraeus, C. Voisset, M. Blondel, Nucleolin directly mediates Epstein-Barr virus immune evasion through binding to G-quadruplexes of EBNA1 mRNA, Nat. Commun. 8 (2017) 16043, https://doi.org/10.1038/ncomms16043.
[12] W.-X. Bian, Y. Xie, X.-N. Wang, G.-H. Xu, B.-S. Fu, S. Li, G. Long, X. Zhou, X.-L. Zhang, Binding of cellular nucleolin with the viral core RNA G-quadruplex structure suppresses HCV replication, Nucleic Acids Res. 47 (2019) 56–68, https://doi.org/10.1093/nar/gky1177.
[13] A.D. Cian, J.-L. Mergny, Quadruplex ligands may act as molecular chaperones for tetramolecular quadruplex formation, Nucleic Acids Res. 35 (2007) 2483–2493, https://doi.org/10.1093/nar/gkm098.
[14] Y. Kusov, J. Tan, E. Alvarez, L. Enjuanes, R. Hilgenfeld, A G-quadruplex-binding macrodomain within the "SARS-unique domain" is essential for the activity of the SARS-coronavirus replication–transcription complex, Virology 484 (2015) 313–322, https://doi.org/10.1016/j.virol.2015.06.016.
[15] C. Platella, C. Riccardi, D. Montesarchio, G.N. Roviello, D. Musumeci, G-quadruplex-based aptamers against protein targets in therapy and diagnostics, Biochim. Biophys. Acta Gen. Subj. 1861 (2017) 1429–1447, https://doi.org/10.1016/j.bbagen.2016.11.027.
[16] J.L. Huppert, S. Balasubramanian, Prevalence of quadruplexes in the human genome, Nucleic Acids Res. 33 (2005) 2908–2916, https://doi.org/10.1093/nar/gki609.
[17] J. Eddy, N. Maizels, Gene function correlates with potential for G4 DNA formation in the human genome, Nucleic Acids Res. 34 (2006) 3887–3896, https://doi.org/10.1093/nar/gkl529.
[18] A. Bedrat, L. Lacroix, J.-L. Mergny, Re-evaluation of G-quadruplex propensity with G4Hunter, Nucleic Acids Res. 44 (2016) 1746–1759, https://doi.org/10.1093/nar/gkw006.
[19] V. Brázda, J. Kolomazník, J. Lýsek, M. Bartas, M. Fojta, J. Šťastný, J.-L. Mergny, G4Hunter web application: a web server for G-quadruplex prediction, Bioinformatics 35 (2019) 3493–3495, https://doi.org/10.1093/bioinformatics/btz087.
[20] B. Ding, Viroids: self-replicating, mobile, and fast-evolving noncoding regulatory RNAs, WIREs RNA 1 (2010) 362–375, https://doi.org/10.1002/wrna.22.
[21] P. Gnanasekaran, S. Chakraborty, Biology of viral satellites and their role in pathogenesis, Curr. Opin. Virol. 33 (2018) 96–105, https://doi.org/10.1016/j.coviro.2018.08.002.
[22] A.M. Fleming, Y. Ding, A. Alenko, C.J. Burrows, Zika virus genomic RNA possesses conserved G-quadruplexes characteristic of the Flaviviridae family, ACS Infect. Dis. 2 (2016) 674–681, https://doi.org/10.1021/acsinfecdis.6b00109.
[23] E. Puig Lombardi, A. Londoño-Vallejo, A. Nicolas, Relationship between G-quadruplex sequence composition in viruses and their hosts, Molecules 24 (2019) 1942, https://doi.org/10.3390/molecules24101942.
[24] E. Lavezzo, M. Berselli, I. Frasson, R. Perrone, G. Palù, A.R. Brazzale, S.N. Richter, S. Toppo, G-quadruplex forming sequences in the genome of all known human viruses: a comprehensive guide, PLoS Comput. Biol. 14 (2018), e1006675, https://doi.org/10.1371/journal.pcbi.1006675.
[25] T. Mihara, Y. Nishimura, Y. Shimizu, H. Nishiyama, G. Yoshikawa, H. Uehara, P. Hingamp, S. Goto, H. Ogata, Linking virus genomes with host taxonomy, Viruses 8 (2016) 66, https://doi.org/10.3390/v8030066.
[26] ROC analysis: web-based calculator for ROC curves. http://www.rad.jhmi.edu/jeng/javarad/roc/JROCFITi.html#exportAnchor. (Accessed 29 March 2021) accessed.
[27] P.J. Walker, S.G. Siddell, E.J. Lefkowitz, A.R. Mushegian, D.M. Dempsey, B.E. Dutilh, B. Harrach, R.L. Harrison, R.C. Hendrickson, S. Junglen, N.J. Knowles, A.M. Kropinski, M. Krupovic, J.H. Kuhn, M. Nibert, L. Rubino, S. Sabanadzovic, P. Simmonds, A. Varsani, F.M. Zerbini, A.J. Davison, Changes to virus taxonomy and the international code of virus classification and nomenclature ratified by the international committee on taxonomy of viruses (2019), Arch. Virol. 164 (2019) 2417–2429, https://doi.org/10.1007/s00705-019-04306-w.
[28] D. Baltimore, Expression of animal virus genomes, Bacteriol. Rev. 35 (1971) 235–241.
[29] N. Philippe, M. Legendre, G. Doutre, Y. Couté, O. Poirot, M. Lescot, D. Arslan, V. Seltzer, L. Bertaux, C. Bruley, J. Garin, J.-M. Claverie, C. Abergel, Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes, Science 341 (2013) 281–286, https://doi.org/10.1126/science.1239181.
[30] S. Spandole, D. Cimponeriu, L.M. Berca, G. Mihăescu, Human anelloviruses: an update of molecular, epidemiological and clinical aspects, Arch. Virol. 160 (2015) 893–908, https://doi.org/10.1007/s00705-015-2363-9.
[31] P.M. Lieberman, Epigenetics and genetics of viral latency, Cell Host Microbe 19 (2016) 619–628, https://doi.org/10.1016/j.chom.2016.04.008.
[32] M. Kane, T. Golovkina, Common threads in persistent viral infections, J. Virol. 84 (2010) 4116–4123, https://doi.org/10.1128/JVI.01905-09.
[33] Coronaviridae, Fenners Vet. Virol. (2011) 393–413, https://doi.org/10.1016/B978-0-12-375158-4.00024-9.
[34] R.M. Buller, G.J. Palumbo, Poxvirus pathogenesis, Microbiol. Rev. 55 (1991) 80–122.
[35] V. Brázda, Y. Luo, M. Bartas, P. Kaura, O. Porubiaková, J. Šťastný, P. Pečinka, D. Verga, V. Da Cunha, T.S. Takahashi, P. Forterre, H. Myllykallio, M. Fojta, J.-L. Mergny, G-quadruplexes in the archaea domain, Biomolecules 10 (2020) 1349, https://doi.org/10.3390/biom10091349.
[36] E.-C. Shin, P.S. Sung, S.-H. Park, Immune responses and immunopathology in acute and chronic viral hepatitis, Nat. Rev. Immunol. 16 (2016) 509–523, https://doi.org/10.1038/nri.2016.69.
[37] C.L. Charlton, E. Babady, C.C. Ginocchio, T.F. Hatchette, R.C. Jerris, Y. Li, M. Loeffelholz, Y.S. McCarter, M.B. Miller, S. Novak-Weekley, A.N. Schuetz, Y.-W. Tang, R. Widen, S.J. Drews, Practical guidance for clinical microbiology laboratories: viruses causing acute respiratory tract infections, Clin. Microbiol. Rev. 32 (2018), https://doi.org/10.1128/CMR.00042-18.

[38] E. van der Vries, K.J. Stittelaar, G. van Amerongen, E.J.B. Veldhuis Kroeze, L. de Waal, P.L.A. Fraaij, R.J. Meesters, T.M. Luider, B. van der Nagel, B. Koch, A.G. Vulto, M. Schutten, A.D.M.E. Osterhaus, Prolonged influenza virus shedding and emergence of antiviral resistance in immunocompromised patients and ferrets, PLoS Pathog. 9 (2013), e1003343, https://doi.org/10.1371/journal.ppat.1003343.
[39] R. Gonzalez-Dosal, K.A. Horan, S.H. Rahbek, H. Ichijo, Z.J. Chen, J.J. Mieyal, R. Hartmann, S.R. Paludan, HSV infection induces production of ROS, which potentiate signaling from pattern recognition receptors: role for S-glutathionylation of TRAF3 and 6, PLoS Pathog. 7 (2011), e1002250, https://doi.org/10.1371/journal.ppat.1002250.
[40] G.A. Maglennon, J. Doorbar, The biology of papillomavirus latency, Open Virol. J. 6 (2012) 190–197, https://doi.org/10.2174/1874357901206010190.
[41] T. Lion, Adenovirus persistence, reactivation, and clinical management, FEBS Lett. 593 (2019) 3571–3582, https://doi.org/10.1002/1873-3468.13576.
[42] J.S. Niczyporuk, Adenoviruses and their diversity in poultry, Appl. Genet. Genomics Poult. Sci. 7 (2018) 103–118, https://doi.org/10.5772/intechopen.77131.
[43] Family - Polyomaviridae, in: A.M.Q. King, M.J. Adams, E.B. Carstens, E.J. Lefkowitz (Eds.), Virus Taxon., Elsevier, San Diego, 2012, pp. 279–290, https://doi.org/10.1016/B978-0-12-384684-6.00027-6.
[44] T.L. Goldberg, V.L. Clyde, A. Gendron-Fitzpatrick, S.D. Sibley, R. Wallace, Severe neurologic disease and chick mortality in crested screamers (Chauna torquata) infected with a novel Gyrovirus, Virology 520 (2018) 111–115, https://doi.org/10.1016/j.virol.2018.05.014.
[45] K. Dhama, R.S. Chauhan, M. Mahendran, S.V.S. Malik, Rotavirus diarrhea in bovines and other domestic animals, Vet. Res. Commun. 33 (2009) 1–23, https://doi.org/10.1007/s11259-008-9070-x.
[46] S. Roy, L.H. Vandenberghe, S. Kryazhimskiy, R. Grant, R. Calcedo, X. Yuan, M. Keough, A. Sandhu, Q. Wang, C.A. Medina-Jaszek, J.B. Plotkin, J.M. Wilson, Isolation and characterization of adenoviruses persistently shed from the gastrointestinal tract of non-human primates, PLoS Pathog. 5 (2009), e1000503, https://doi.org/10.1371/journal.ppat.1000503.
[47] F. Krammer, G.J.D. Smith, R.A.M. Fouchier, M. Peiris, K. Kedzierska, P.C. Doherty, P. Palese, M.L. Shaw, J. Treanor, R.G. Webster, A. García-Sastre, Influenza, Nat. Rev. Dis. Primer. 4 (2018) 3, https://doi.org/10.1038/s41572-018-0002-y.
[48] F.B. Lestari, S. Vongpunsawad, N. Wanlapakorn, Y. Poovorawan, Rotavirus infection in children in Southeast Asia 2008–2018: disease burden, genotype distribution, seasonality, and vaccination, J. Biomed. Sci. 27 (2020) 66, https://doi.org/10.1186/s12929-020-00649-8.
[49] M. Weidner-Glunde, E. Kruminis-Kaszkiel, M. Savanagouder, Herpesviral latency—common themes, Pathogens 9 (2020) 125, https://doi.org/10.3390/pathogens9020125.
[50] K. Murray, C. Walker, E. Herrington, J.A. Lewis, J. McCormick, D.W.C. Beasley, R.B. Tesh, S. Fisher-Hoch, Persistent infection with west nile virus years after initial infection, J. Infect. Dis. 201 (2010) 2–4, https://doi.org/10.1086/648731.
[51] L. Mlera, W. Melik, M.E. Bloom, The role of viral persistence in flavivirus biology, Pathog. Dis. 71 (2014) 137–163, https://doi.org/10.1111/2049-632X.12178.
[52] Family - Reoviridae, in: A.M.Q. King, M.J. Adams, E.B. Carstens, E.J. Lefkowitz (Eds.), Virus Taxon., Elsevier, San Diego, 2012, pp. 541–637, https://doi.org/10.1016/B978-0-12-384684-6.00051-3.
[53] A. Mrzljak, I. Tabain, H. Premac, M. Bogdanic, L. Barbic, V. Savic, V. Stevanovic, A. Jelic, D. Mikulic, T. Vilibic-Cavlek, The role of emerging and neglected viruses in the etiology of hepatitis, Curr. Infect. Dis. Rep. 21 (2019) 51, https://doi.org/10.1007/s11908-019-0709-2.
[54] F.A. Lempp, S. Urban, Hepatitis delta virus: replication strategy and upcoming therapeutic options for a neglected human pathogen, Viruses 9 (2017) 172, https://doi.org/10.3390/v9070172.
[55] T. Horvatits, J. Schulze zur Wiesch, M. Lütgehetmann, A.W. Lohse, S. Pischke, The clinical perspective on hepatitis E, Viruses 11 (2019) 617, https://doi.org/10.3390/v11070617.
[56] H.L. Lightfoot, T. Hagen, N.J. Tatum, J. Hall, The diverse structural landscape of quadruplexes, FEBS Lett. 593 (2019) 2083–2102, https://doi.org/10.1002/1873-3468.13547.
[57] G.P. Martelli, S. Sabanadzovic, N. Abou Ghanem-Sabanadzovic, P. Saldarelli, Maculavirus, a new genus of plant viruses, Arch. Virol. 147 (2002) 1847–1853, https://doi.org/10.1007/s007050200046.
[58] R. Perrone, M. Nadai, I. Frasson, J.A. Poe, E. Butovskaya, T.E. Smithgall, M. Palumbo, G. Palù, S.N. Richter, A dynamic G-quadruplex region regulates the HIV-1 long terminal repeat promoter, J. Med. Chem. 56 (2013) 6521–6530, https://doi.org/10.1021/jm400914r.
[59] A.M. Fleming, N.L.B. Nguyen, C.J. Burrows, Colocalization of m6A and G-quadruplex-forming sequences in viral RNA (HIV, Zika, hepatitis B, and SV40) suggests topological control of adenosine N6-methylation, ACS Cent. Sci. 5 (2019) 218–228, https://doi.org/10.1021/acscentsci.8b00963.
[60] R. Perrone, M. Nadai, J.A. Poe, I. Frasson, M. Palumbo, G. Palù, T.E. Smithgall, S.N. Richter, Formation of a unique cluster of G-quadruplex structures in the HIV-1 Nef coding region: implications for antiviral activity, PloS One 8 (2013), e73121, https://doi.org/10.1371/journal.pone.0073121.
[61] B. Biswas, P. Kumari, P. Vivekanandan, Pac1 signals of human Herpesviruses contain a highly conserved G-quadruplex motif, ACS Infect. Dis. 4 (2018) 744–751, https://doi.org/10.1021/acsinfecdis.7b00279.
[62] S. Artusi, M. Nadai, R. Perrone, M.A. Biasolo, G. Palù, L. Flamand, A. Calistri, S.N. Richter, The Herpes Simplex Virus-1 genome contains multiple clusters of repeated G-quadruplex: implications for the antiviral activity of a G-quadruplex ligand, Antivir. Res. 118 (2015) 123–131, https://doi.org/10.1016/j.antiviral.2015.03.016.
[63] H. Deng, B. Gong, Z. Yang, Z. Li, H. Zhou, Y. Zhang, X. Niu, S. Liu, D. Wei, Intensive distribution of G2-quaduplexes in the pseudorabies virus genome and their sensitivity to cations and G-quadruplex ligands, Molecules 24 (2019) 774, https://doi.org/10.3390/molecules24040774.
[64] S. Kumar, D. Choudhary, A. Patra, N.S. Bhavesh, P. Vivekanandan, Analysis of G-quadruplexes upstream of herpesvirus miRNAs: evidence of G-quadruplex mediated regulation of KSHV miR-K12-1-9,11 cluster and HCMV miR-US33, BMC Mol. Cell Biol. 21 (2020) 67, https://doi.org/10.1186/s12860-020-00306-w.
[65] P. Murat, J. Zhong, L. Lekieffre, N.P. Cowieson, J.L. Clancy, T. Preiss, S. Balasubramanian, R. Khanna, J. Tellam, G-quadruplexes regulate Epstein-Barr virus–encoded nuclear antigen 1 mRNA translation, Nat. Chem. Biol. 10 (2014) 358–364, https://doi.org/10.1038/nchembio.1479.
[66] P. Majee, U. Shankar, S. Pasadi, K. Muniyappa, D. Nayak, A. Kumar, Genome-wide analysis reveals a regulatory role for G-quadruplexes during Adenovirus multiplication, Virus Res. (2020) 197960, https://doi.org/10.1016/j.virusres.2020.197960.
[67] B. Biswas, M. Kandpal, U.K. Jauhari, P. Vivekanandan, Genome-wide analysis of G-quadruplexes in herpesvirus genomes, BMC Genom. 17 (2016) 949, https://doi.org/10.1186/s12864-016-3282-1.
[68] K. Tlučková, M. Marušič, P. Tóthová, L. Bauer, P. Šket, J. Plavec, V. Viglasky, Human papillomavirus G-quadruplexes, Biochemistry 52 (2013) 7207–7216, https://doi.org/10.1021/bi400897g.
[69] N. Bohálová, A. Cantara, M. Bartas, P. Kaura, J. Šťastný, P. Pečinka, M. Fojta, V. Brázda, Tracing dsDNA virus–host coevolution through correlation of their G-quadruplex-forming sequences, Int. J. Mol. Sci. 22 (2021) 3433, https://doi.org/10.3390/ijms22073433.
[70] C. Jaubert, A. Bedrat, L. Bartolucci, C. Di Primo, M. Ventura, J.-L. Mergny, S. Amrane, M.-L. Andreola, RNA synthesis is modulated by G-quadruplex formation in Hepatitis C virus negative RNA strand, Sci. Rep. 8 (2018) 8120, https://doi.org/10.1038/s41598-018-26582-3.
[71] S.T. Jacob, I. Crozier, W.A. Fischer, A. Hewlett, C.S. Kraft, M.-A. de L. Vega, M.J. Soka, V. Wahl, A. Griffiths, L. Bollinger, J.H. Kuhn, Ebola virus disease, Nat. Rev. Dis. Primer. 6 (2020) 13, https://doi.org/10.1038/s41572-020-0147-3.
[72] K.M. Coffin, J. Liu, T.K. Warren, C.D. Blancett, K.A. Kuehl, D.K. Nichols, J.J. Bearss, C.W. Schellhase, C.J. Retterer, J.M. Weidner, S.R. Radoshitzky, J.M. Brannan, A.P. Cardile, J.M. Dye, G. Palacios, M.G. Sun, J.H. Kuhn, S. Bavari, X. Zeng, Persistent Marburg virus infection in the testes of nonhuman primate survivors, Cell Host Microbe 24 (2018) 405–416, https://doi.org/10.1016/j.chom.2018.08.003, e3.
[73] S.-R. Wang, Q.-Y. Zhang, J.-Q. Wang, X.-Y. Ge, Y.-Y. Song, Y.-F. Wang, X.-D. Li, B.-S. Fu, G.-H. Xu, B. Shu, P. Gong, B. Zhang, T. Tian, X. Zhou, Chemical targeting of a G-quadruplex RNA in the Ebola virus L gene, Cell Chem. Biol. 23 (2016) 1113–1122, https://doi.org/10.1016/j.chembiol.2016.07.019.
[74] P. Krafčíková, E. Demkovičová, V. Víglaský, Ebola virus derived G-quadruplexes: thiazole orange interaction, Biochim. Biophys. Acta BBA - Gen. Subj. 1861 (2017) 1321–1328, https://doi.org/10.1016/j.bbagen.2016.12.009.
[75] D.E. Griffin, W.-H.W. Lin, A.N. Nelson, Understanding the causes and consequences of measles virus persistence, F1000Research 7 (2018) 237, https://doi.org/10.12688/f1000research.12094.1.
[76] L. Hároníková, J. Coufal, I. Kejnovská, E.B. Jagelská, M. Fojta, P. Dvořáková, P. Muller, B. Vojtesek, V. Brázda, IFI16 preferentially binds to DNA with quadruplex structure and enhances DNA quadruplex formation, PloS One 11 (2016), e0157156, https://doi.org/10.1371/journal.pone.0157156.
[77] G. Pisano, A. Roy, M. Ahmed Ansari, B. Kumar, L. Chikoti, B. Chandran, Interferon-γ-inducible protein 16 (IFI16) is required for the maintenance of Epstein-Barr virus latency, Virol. J. 14 (2017) 221, https://doi.org/10.1186/s12985-017-0891-5.
[78] V. Brázda, O. Porubiaková, A. Cantara, N. Bohálová, J. Coufal, M. Bartas, M. Fojta, J.-L. Mergny, G-quadruplexes in H1N1 influenza genomes, BMC Genom. 22 (2021) 77, https://doi.org/10.1186/s12864-021-07377-9.
[79] M. Bartas, V. Brázda, N. Bohálová, A. Cantara, A. Volná, T. Stachurová, K. Malachová, E.B. Jagelská, O. Porubiaková, J. Červeň, P. Pečinka, In-depth bioinformatic analyses of nidovirales including human SARS-CoV-2, SARS-CoV, MERS-CoV viruses suggest important roles of non-canonical nucleic acid structures in their lifecycles, Front. Microbiol. 11 (2020) 1583, https://doi.org/10.3389/fmicb.2020.01583.
[80] D. Ji, M. Juhas, C.M. Tsang, C.K. Kwok, Y. Li, Y. Zhang, Discovery of G-quadruplex-forming sequences in SARS-CoV-2, brief, Bioinformation 22 (2021) 1150–1160, https://doi.org/10.1093/bib/bbaa114.
[81] D. Piekna-Przybylska, G. Sharma, S.B. Maggirwar, R.A. Bambara, Deficiency in DNA damage response, a new characteristic of cells infected with latent HIV-1, Cell Cycle 16 (2017) 968–978, https://doi.org/10.1080/

15384101.2017.1312225.
[82] J.T. Tellam, J. Zhong, L. Lekieffre, P. Bhat, M. Martinez, N.P. Croft, W. Kaplan, R.L. Tellam, R. Khanna, mRNA Structural constraints on EBNA1 synthesis impact on in vivo antigen presentation and early priming of CD8+ T cells, PLoS Pathog. 10 (2014), e1004423, https://doi.org/10.1371/journal.ppat.1004423.
[83] P. Dabral, J. Babu, A. Zareie, S.C. Verma, LANA and hnRNP A1 regulate the translation of LANA mRNA through G-quadruplexes, J. Virol. 94 (2019) e01508–e01519, https://doi.org/10.1128/JVI.01508-19.